Information page for written examinations at Linköping University

Examination date	2017-05-10
Room (1)	TER3(4)
Time	8-12
Course code	732A54
Exam code	TENT
Course name **Exam name**	Big Data Analytics (Analys av Big data) Examination (Tentamen)
Department	IDA
Number of questions in the examination	14
Teacher responsible/contact person during the exam time	Christoph Kessler, Jose Pena, Valentina Ivanova (Patrick Lambrix)
Contact number during the exam time	2406 / 1651 / 2605
Visit to the examination room approximately	10:00
Name and contact details to the course administrator (name + phone nr + mail)	Annelie Almquist, 2934, annelie.almquist@liu.se
Equipment permitted	dictionary
Other important information	
Number of exams in the bag	

Exam

732A54 Big Data Analytics

May 10, 2017, 8-12

Grades: For a pass grade you need to obtain 50% of the total points.

Instructions:

In addition to the instructions on the cover page:

- Write clearly.
- Start the answers to a question on a new page.
- If you make assumptions that are not given in a question, then clearly describe these assumptions. (Of course, these assumptions cannot change the exercise.)
- Give relevant answers to the questions. Points can be deducted for answers that are not answers to the question.
- Answer in English.

Question 1 (2p)

Compare RDBMS and NoSQL data management systems regarding:
- data models
- consistency and availability in the presence of network partitions.

Question 2 (3p)

Explain the role of name nodes and data nodes in the HDFS.

Question 3 (2p)

P1, P2 and P3 are three distributed processes. The events following below have occurred during the processes and the values for their vector clocks are given:

```
P1: A (0,0,0); B (1,0,0); C (2,0,0); D (3,0,0); E (4,0,2)
P2: F (0,0,0); G (1,1,0); H (2,2,0); I (2,3,3)
P3: J (0,0,0); K (0,0,1); L (0,0,2); M (0,0,3)
```

Name the relationships between the following two pairs of events and explain (with the help of the respective formal rules) how you have determined them:
- B (1,0,0) and K (0,0,1)
- I (2,3,3) and E (4,0,2).

Question 4 (3p)

Explain the HBase data model.

Question 5 (1+1=2p)

(a) Explain (including an annotated drawing) the principle of memory hierarchy as used in modern server computers.
(b) What is the purpose of memory hierarchy in computer architecture? What kind of programs are expected to benefit from memory hierarchy, and why?

Question 6 (1p)

Give an example of a network (name and short explanation including an annotated drawing) that is not scalable (i.e., its accumulated throughput does not grow with the number of processors).

Question 7 (0.5p)

Define the parallel algorithmic design pattern *data parallelism.*

Question 8 (1p)

What properties do functions need to fulfill that are to be used in Combine or Reduce steps of MapReduce, and why?

Question 9 (1.5p)

Which steps of MapReduce involve disk I/O, and for what purpose?

Question 10 (1.5+0.5=2p)

(a) Shortly explain and compare the fault tolerance mechanisms of MapReduce and Spark.
(b) In particular, which of the two is expected to result in less disk I/O if the probability of hardware failures is relatively low? Explain your answer.

Question 11 (1p)

For iterative machine learning jobs, Spark usually outperforms MapReduce by an order of magnitude. Explain why.

Question 12 (1p)

What do systems such as YARN and Mesos do?

Question 13 (6p)

Implement in Spark (PySpark) the following k-means algorithm.

```
1  Assign each point to a cluster at random
2  Compute the cluster centroids as the averages of the points assigned to each cluster
3  Repeat the following lines l times
4     Assign each point to the cluster with the closest centroid
5     Update the cluster centroids as the averages of the points assigned to each cluster
```

You can use the functions `randint(A,B)` which produces a random integer in the given interval, and `distance(A,B)` which returns the distance between two points.

Question 14 (2+2=4p)

(a) Describe briefly the main difference between the Spark and MapReduce frameworks. Show where you make use of this distinguishing feature in your implementation of the k-means algorithm.
(b) Give two reasons why the MapReduce framework is suitable for implementing many machine learning algorithms.